

Семинар 2

Задача 1

T1 Дана выборка x_1, \dots, x_n из неизвестного абсолютного непрерывного распределения. Для оценивания плотности в точке 0 используется оценка

$$\hat{p}_n(0) = \frac{\#\{i : x_i \in (-h/2, h/2]\}}{nh}.$$

(i) Докажите, что

$$\begin{aligned} \text{MSE}(\hat{p}_n(0)) &:= \mathbb{E}[(\hat{p}_n(0) - p(0))^2] \\ &= \left(\frac{(p''(0))^2}{576} h^4 + \frac{p(0)}{nh} \right) (1 + o(1)), \quad n \rightarrow \infty, h \rightarrow 0. \end{aligned}$$

Запишем формулу bias-variance tradeoff:

$$\begin{aligned} \text{MSE}(\hat{p}_n(0)) &= \mathbb{E}[(\hat{p}_n(0) - p(0))^2] = \text{Var}(\hat{p}_n(0)) + \text{Bias}^2(\hat{p}_n(0)) \\ \hat{p}_n(0) &= \frac{\#\{i : x_i \in (-h/2, h/2]\}}{nh} = \frac{1}{nh} \sum_{i=1}^n \mathbb{1}\{-h/2 < x_i \leq h/2\} \\ \xi_i &:= \mathbb{1}\{-h/2 < x_i \leq h/2\} = \begin{cases} 1, & \text{if } -h/2 < x_i \leq h/2 \\ 0, & \text{otherwise} \end{cases} \Rightarrow \xi_1, \xi_2, \dots, \xi_n \sim \text{iid } \mathcal{B}(1, p_h) \\ &\Rightarrow \sum_{i=1}^n \mathbb{1}\{-h/2 < x_i \leq h/2\} \sim \mathcal{B}(n, p_h) \\ \mathbb{E}[\hat{p}_n(0)] &= \mathbb{E}\left[\frac{1}{nh} \sum_{i=1}^n \mathbb{1}\{-h/2 < x_i \leq h/2\}\right] = \frac{1}{nh} \mathbb{E}\left[\sum_{i=1}^n \xi_i\right] \stackrel{\text{iid}}{=} \frac{1}{nh} \cdot n \mathbb{E}[\xi_1] = \frac{1}{h} \mathbb{P}\{-h/2 < x_1 \leq h/2\} \\ p_h &= \mathbb{P}\{-h/2 < x_1 \leq h/2\} = \int_{-h/2}^{h/2} p(x) dx = \int_{-h/2}^{h/2} \left(p(0) + p'(0)x + \frac{1}{2} p''(0)x^2 + \bar{o}(x^2) \right) dx \\ &= p(0)h + 0 + \frac{1}{2} p''(0) \cdot \frac{1}{3} \left(\frac{h^3}{8} + \frac{h^3}{8} \right) + \bar{o}(h^3) \\ &= p(0)h + p''(0) \cdot \frac{h^3}{24} + \bar{o}(h^3) \\ \mathbb{E}[\hat{p}_n(0)] &= \frac{1}{h} \left(p(0)h + p''(0) \cdot \frac{h^3}{24} + \bar{o}(h^3) \right) = p(0) + p''(0) \cdot \frac{h^2}{24} + \bar{o}(h^2) \\ \Rightarrow \text{Bias}^2(\hat{p}_n(0)) &= \left(p''(0) \cdot \frac{h^2}{24} + \bar{o}(h^2) \right)^2 = \frac{h^4 (p''(0))^2}{24^2} (1 + \bar{o}(1)) \end{aligned}$$

$$\begin{aligned}
 \bullet \text{Var}(\hat{p}_n(0)) &= \text{Var}\left(\frac{1}{nh} \sum_{i=1}^n \xi_i\right) \stackrel{\text{iid}}{=} \frac{1}{n^2 h^2} \cdot n \text{Var}(\xi_1) = \frac{1}{nh^2} p_h(1-p_h) \\
 &= \frac{1}{nh^2} \left(p(0)h + p''(0)\frac{h^3}{24} + \bar{o}(h^3) \right) \left(1 - p(0)h - p''(0)\frac{h^3}{24} - \bar{o}(h^3) \right) \\
 &= \frac{1}{n} \left(\underbrace{\frac{p(0)}{h}}_{\rightarrow 0, h \rightarrow 0} + \underbrace{p''(0) \cdot \frac{h}{24}}_{\rightarrow 0} + \underbrace{h \bar{o}(1)}_{\rightarrow 0} \right) \left(\underbrace{1 - p(0)h}_{\rightarrow 0} - \underbrace{p''(0) \frac{h^3}{24}}_{\rightarrow 0} + \underbrace{\bar{o}(h^3)}_{\rightarrow 0} \right) \\
 &= \frac{1}{nh} p(0)(1 + \bar{o}(1)), \quad h \rightarrow 0
 \end{aligned}$$

$$\text{MSE}(\hat{p}_n(0)) = \text{Var}(\hat{p}_n(0)) + \text{Bias}^2(\hat{p}_n(0)) = \left(\frac{p(0)}{nh} + \frac{h^4 (p''(0))^2}{576} \right) (1 + \bar{o}(1)) \quad h \rightarrow 0$$

(ii) Является ли $\hat{p}_n(0)$ несмещённой оценкой $p(0)$? Является ли $\hat{p}_n(0)$ состоятельной оценкой $p(0)$?

$$\text{Bias}(\hat{p}_n(0)) = \frac{p''(0)h^2}{24} + \bar{o}(h^2) \neq 0 \Rightarrow \text{смещ.}, \text{ но асимпт. не см. при } h \rightarrow 0$$

$$\begin{aligned} \text{Var}(\hat{p}_n(0)) &= \frac{p(0)}{nh} (1 + \bar{o}(1)) \rightarrow 0 \Leftrightarrow \underline{nh \rightarrow \infty} \Rightarrow \text{ист. при } nh \rightarrow \infty \\ \mathbb{E}[(\hat{p}_n(0) - \mathbb{E}[\hat{p}_n(0)])^2] &\rightarrow 0 \end{aligned}$$

Задача 2

T2 Допустим, что дана выборка из нормального распределения с нулевым средним и дисперсией σ^2 . Вычислите значение параметра bandwidth, минимизирующее AMISE (asymptotic mean integrated squared error) ядерной оценки плотности, построенной на основе гауссовского ядра

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

$$\text{KDE: } \hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

$$\text{AMISE}(\hat{p}_n(x)) = \frac{h^4}{4} \int_{\mathbb{R}} (p''(x))^2 dx \left(\int_{\mathbb{R}} x^2 K(x) dx \right)^4 + \frac{1}{nh} \int_{\mathbb{R}} K^2(x) dx$$

$$\frac{\partial \text{AMISE}(\hat{p}_n(x))}{\partial h} = 0$$

$$h_{\text{opt}} = \frac{1}{n^{1/5}} \left(\frac{\int_{\mathbb{R}} K^2(x) dx}{\int_{\mathbb{R}} (p''(x))^2 dx \left(\int_{\mathbb{R}} x^2 K(x) dx \right)^4} \right)^{1/5}$$

$$\bullet \int_{\mathbb{R}} K^2(x) dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}}$$

$$\bullet \left(\int_{\mathbb{R}} x^2 K(x) dx \right)^4 = \left(\int_{\mathbb{R}} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \right)^4 = \left[\xi \sim N(0,1) \right]^4 = (E\xi^2)^4 = 1^4 = 1$$

$$\begin{aligned} \bullet p''(x) &= \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \right)'' = \left(-\frac{x}{\sqrt{2\pi}\sigma^3} e^{-\frac{x^2}{2\sigma^2}} \right)' = -\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma^3} + \frac{x^2}{\sqrt{2\pi}\sigma^5} e^{-\frac{x^2}{2\sigma^2}} \\ \Rightarrow (p''(x))^2 &= \frac{e^{-x^2/\sigma^2}}{2\pi\sigma^6} - \frac{2x^2}{2\pi\sigma^8} e^{-x^2/\sigma^2} + \frac{x^4}{2\pi\sigma^{10}} e^{-x^2/\sigma^2} \\ \int_{\mathbb{R}} (p''(x))^2 dx &= \int_{\mathbb{R}} \frac{e^{-x^2/\sigma^2}}{2\pi\sigma^6} dx - 2 \int_{\mathbb{R}} \frac{x^2}{2\pi\sigma^8} e^{-x^2/\sigma^2} dx + \int_{\mathbb{R}} \frac{x^4}{2\pi\sigma^{10}} e^{-x^2/\sigma^2} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma^5} \underbrace{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{2}}} e^{-\frac{x^2}{2 \cdot \frac{\sigma^2}{2}}} dx}_{\int_{\mathbb{R}} \phi_{N(0, \sigma^2/2)} = 1} - \frac{2}{\sqrt{2\pi}\sigma^7} \underbrace{\int_{\mathbb{R}} \frac{x^2}{\sqrt{2\pi}\frac{\sigma}{\sqrt{2}}} e^{-\frac{x^2}{2 \cdot \frac{\sigma^2}{2}}} dx}_{\substack{E[x^2], X \sim N(0, \sigma^2/2) \\ x^2 \frac{\sigma}{\sqrt{2}} \xi \Rightarrow E[x^2] = \frac{\sigma^2}{2} E\xi^2 = \frac{\sigma^2}{2} \cdot \frac{1}{2} = \frac{\sigma^2}{4}}} + \frac{1}{\sqrt{2\pi}\sigma^9} \underbrace{\int_{\mathbb{R}} \frac{x^4}{\sqrt{2\pi}\frac{\sigma}{\sqrt{2}}} e^{-\frac{x^2}{2 \cdot \frac{\sigma^2}{2}}} dx}_{\substack{E[x^4], \dots \\ = \frac{\sigma^4}{4} E\xi^4 = \frac{3\sigma^4}{4}}} \\ &= \frac{1}{\sqrt{2\pi}\sigma^5} - \frac{1}{\sqrt{2\pi}\sigma^5} + \frac{3}{8\sqrt{2\pi}\sigma^5} = \frac{3}{8\sqrt{2\pi}\sigma^5} \\ \Rightarrow I_{\text{opt}} &= \frac{1}{n^{1/5}} \left(\frac{8\sqrt{2\pi}\sigma^5}{3} \right)^{1/5} = \frac{\sigma}{n^{1/5}} \left(\frac{4}{3} \right)^{1/5} \rightarrow \text{"nrd" in } \mathbb{R} \end{aligned}$$

Задача 3

ТЗ Пусть ψ_0, ψ_1, \dots - ортонормированный базис Лежандра в пространстве $L^2([-1, 1])$. Докажите, что функция

$$K(x) = \sum_{m=0}^n \psi_m(0) \psi_m(x) \mathbb{I}\{|x| \leq 1\}$$

является ядром порядка n в том смысле, что $\int_{\mathbb{R}} K(x) dx = 1$ и $\int_{\mathbb{R}} K(x) x^k dx = 0$ для $k = 1, \dots, n$. $\rightarrow \int_{\mathbb{R}} x^k K(x) dx, k \in [0, n]$

$$\int_{-1}^1 \psi_i(x) \psi_j(x) dx = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

$x^k \in L^2([-1, 1])$, $x^k = \sum_{i=0}^k \psi_i(x) \alpha_{ik}$ $\rightarrow \psi_k$ имеет степень k

$$\int_{\mathbb{R}} x^k K(x) dx = \int_{-1}^1 \sum_{i=0}^k \psi_i(x) \alpha_{ik} \sum_{m=0}^n \psi_m(0) \psi_m(x) dx = \sum_{i=0}^k \sum_{m=0}^n \alpha_{ik} \psi_m(0) \underbrace{\int_{-1}^1 \psi_i(x) \psi_m(x) dx}_{=0, i \neq m}$$

$$= \sum_{i=0}^{\min\{k, n\}} \underbrace{\alpha_{ik} \psi_m(0)}_{x^k|_{x=0}} \cdot \underbrace{\int_{-1}^1 \psi_m^2(x) dx}_1 = \begin{cases} 0, & k \in [1, n] \\ 1, & k = 0 \end{cases}$$

Задача 4

Т4 Оценка кросс-валидации параметра h определяется как точка минимума функции

$$\hat{J}(h) = \int \hat{p}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{p}_{(-i)}(x_i),$$

где $\hat{p}_{(-i)}$ - оценка, построенная по всем значениям, кроме i -го (leave-one-out estimate). Для ядерной оценки плотности \hat{p}_n , представьте функцию $\hat{J}(h)$ в виде функции, зависящей от выборки x_1, \dots, x_n только через разности $x_i - x_j$, $i, j = 1, \dots, n$.

$$\int_{\mathbb{R}} (\hat{p}_n(x) - p(x))^2 dx = \int_{\mathbb{R}} \hat{p}_n^2(x) dx - 2 \int_{\mathbb{R}} \hat{p}_n(x) p(x) dx + \int_{\mathbb{R}} p^2(x) dx \rightarrow \min$$

$$\begin{aligned} \hat{J}(h) &= \int_{\mathbb{R}} \frac{1}{n^2 h^2} \left(\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \right)^2 dx - \frac{2}{n} \sum_{i=1}^n \sum_{j \neq i} \frac{1}{(n-1)h} K\left(\frac{x_i - x_j}{h}\right) \\ &= \frac{2}{n} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{(n-1)h} K\left(\frac{x_i - x_j}{h}\right) - \frac{2}{n} \cdot n \cdot \frac{1}{(n-1)h} K(0) \\ &= \frac{1}{n^2 h^2} \int_{\mathbb{R}} \sum_{i=1}^n \sum_{j=1}^n K\left(\frac{x - x_i}{h}\right) K\left(\frac{x - x_j}{h}\right) dx - \frac{2}{n^2 h^2} \int_{\mathbb{R}} \sum_{i=1}^n \sum_{j=1}^n K(u) K\left(u + \frac{x_i - x_j}{h}\right) h du \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n \sum_{j=1}^n \int_{\mathbb{R}} K(u) K\left(\frac{x_i - x_j}{h} - u\right) h du \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n K \star K\left(\frac{x_i - x_j}{h}\right) \end{aligned}$$

$$\hat{J}(h) = \frac{1}{n^2 h} \sum_{i,j=1}^n K \star K\left(\frac{x_i - x_j}{h}\right) - \frac{2}{n} \sum_{j=1}^n \frac{1}{(n-1)h} K\left(\frac{x_j - x_j}{h}\right) + \frac{dn}{(n-1)h} K(0)$$

$$h = \underset{h > 0}{\operatorname{argmin}} \hat{J}(h) \rightarrow \text{"new"}$$

$$\text{"bw": } \underset{h > 0}{\operatorname{argmin}} \left\{ \frac{1}{n^2 h} \sum_{i,j=1}^n (K \star K\left(\frac{x_i - x_j}{h}\right) - 2K\left(\frac{x_j - x_j}{h}\right)) + \frac{dn}{(n-1)h} K(0) \right\}$$